HEB

CASS

A Revisit to Clustering Techniques with its Application in Agriculture Sector

**Vyoma Srivastava, **Dr.K.K Aggarwal & ***Dr. Abhay Kumar Srivastava*

*Dept. of Computer Science, Jaipur National University, Jaipur, India

**Dept. of Computer Science, Jaipur National University, Jaipur, India

***Amity Business School, Amity University, Noida, India

Address for Correspondence: serviceheb@gmail.com

ABSTRACT

Inspite of the big challenge of missing data for data mining algorithms, Data mining has made a great progress in recent years. Data Mining is a process where its techniques are used to extract some useful knowledge from a large data base. It is very helpful since it helps human race discover knowledge out of data and presenting it in a form that is easily understood. Data mining uses various methods and techniques which allow to analyze very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. [1]

Today, where Data Mining has found its application in various Industries, We believe that Data Mining techniques like Agglomerative Clustering, DBSCAN, EM Algorithms, K-Means applications can bring a good advancement in agricultural sector also. A comprehensive review of the applications of clustering techniques in agriculture domain is presented in this paper.

KEYWORDS:

Data mining, clustering, classification of clustering, Mixed Attributes, Algorithms

Access this Article Online	
http://heb-nic.in/cass-studies	**Quick Response Code:**
Received on 25/03/2019 Accepted on 11/04/2019@HEB All rights reserved	

QUICK GLANCE COMPARISON OF CLUSTERING TECHNIQUES[2]:

Name	Algorithm	Key-Idea	Type of Data	Advantages	Disadvantage
Partitional	K-means	Mean Centroid	Numerical	Simple Most Popular	Sensitive to outliers Centroids not meaningful in most Problems
	PAM	Mediod -centriod		robust to outliers	Cluster should be pre-determined
	CLARA			Applicable for large data set	Sensitive to outliers
	CLARANS			Handles outliers effectively	High Cost
Density Based	DBSCAN	Fixed size	Numerical	Resistant to noise Can handle clusters of various shapes and sizes	Cannot handle varying densities
	OPTICS	Variable size		Good for data set with large amount of noise Faster in computation	Needs large no.of parameters
	DENCLUE			Solid mathematical foundation	Needs large no.of parameters
	RDBC			More effective in discovering varied shape clusters Handles noise effectively	Cost Varying
Hierarchical agglomerative	CURE	Partition Samples	Numerical	Robust to outliers Appropriate for handling large dataset	Ignores information about inter-connectivity of objects
	BIRCH	Multidimensional	Numerical	suitable for large databases scales linearly	Handles only numeric data sensitive to data records
	ROCK	Notion of Links	Categorical	Robust Appropriate for large dataset	space complexity depends on initialization of local heaps
	S-Link	Closest pair of points	-	it does not need to specify no.of clusters	Termination condition needs to be Satisfied Sensitive to outliers
	Ave-Link	Centriod of clusters	-	It considers all members in cluster rather than single point	It produces clusters with same Variance
	Com-Link	Farthest pair of points	-	Not strongly affected by outliers	It has problem with convex shape Clusters
Grid	STING	Multiple Grids	Numerical	Allows parallelization and multiresolution	Does not define appropriate level of Granularity
	WaveClusters		Numerical	High-quality clusters Successful outlier handling	Cost Varying
	CLIQUE	Density based grids		Dimensionality reduction Scalability Insensitive to noise	Prone to high dimensional clusters

Data mining tasks can be classified into two categories:

- Predictive
- Descriptive

Predictive data mining is used to predict the direct values based on patterns determined from known results while Descriptive data mining establish and elaborate the general properties of the data in the database. Prediction uses few fields or variables of the database to predict the future values of other pertinent variables. Predictive data mining approach is more commonly used across the board. Extrapolations done by using Predictive data mining techniques can help in agriculture like predicting future crops, effects of specific fertilizer or pesticide, Weather forecasting, revenue prediction etc. etc.[3]

In this paper we humbly present a review of some Clustering technique on the basis of the algorithms. The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. [4] It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can be done by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multi-resolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

CLUSTERING METHODOLOGIES

HIERARCHICAL CLUSTERING / CONNECTIVITY BASED CLUSTERING:

Popularly known as HCA or Hierarchical Clustering Analysis is a method which builds a hierarchy of clusters and is either Agglomerative or Divisive. Whereas Divisive Hierarchical Clustering uses the Top-Down approach, the Agglomerative Hierarchical Clustering uses Bottom-Up approach, 0020 where, each data point is put into a cluster and pairs of clusters keep getting merged aswe move up the Hierarchy. In HAC (Hierarchical Agglomerative Clustering), usage of Distance matrix is very important, using which the Distance between the elements is determined and the elements closest to each other are generally merged into the cluster. The merger continues from bottom till the top based on updated distance matrix with each merger. DIANA is the basic principle behind Divisive Clustering [5]. ‘DI’visive‘ ANA’lysis clustering Algorithm starts with whole dataset as a single cluster and splits the cluster based on maximum average dissimilarity, while moving bottom from the top.

CENTROID BASED CLUSTERING / K-MEANS CLUSTERING:

Clusters are represented by central vector, which need not be a dataset in itself. When the numbers of clusters is fixed to k, k-means clustering gives an optimized solution. One of the most known ways of K-Means Clustering is Lloyd’s Algorithm [6], where a local optimum is found and is run multiple times with different random clusters. Variation of k-means include these optimizations as choosing the best of the multiple runs. It also restricts the central vector to members of data sets (kmedoids), selecting medians (k-medians clustering) or a fuzzy cluster assignment (fuzzy c-means) or in choosing initial centers less randomly (k-means++). All the algorithms demand k i.e. number of clusters in advance and of approximate similar size since they assign an object to nearest centroid. This may lead to cutting of borders which is not a big issue since this algorithm optimizes cluster center and not borders. It may be seen as a model-based clustering, and Lloyd's algorithm as variation of Expectation-Maximization algorithm. It is closest to nearest neighbour classification conceptually, which is also used a lot in machine learning. This method partitions the dataset into Voronoi diagrams.

DISTRIBUTION BASED CLUSTERING / GAUSSIAN MIXTURE MODEL CLUSTERING:

This method uses Expectation-Maximization Algorithm for Gaussian Mixture model. Dataset is modelled with fixed number of Gaussian distributions, to avoid overfitting. Gaussian distributions used are the ones which are initialized randomly and parameters of whose are optimized post multiple iterations to better fit the dataset. All this converges to a local optimum, which may be different with every run. To obtain a hard clustering, objects are assigned to their respective Gaussian distribution. Though the same may not needed for soft clustering. This clustering produces complex results, which depict dependence and correlation between the attributes. Gaussian distribution is one of the most common continuous probability distributions. Expectation-Maximization Algorithm [7] is a method involving multiple iterations to find maximum likelihood estimates or parameters in statistical models.

DENSITY BASED CLUSTERING:

Clusters, in this method, are defined as areas of higher density, as compared to the other members of the dataset [8].DBSCAN [9] happens to be most known density based clustering method. It follows the logic of "density reachability", which is based on distance threshold being used as the connection methodology. DBSCAN is a non-complex method which needs linear number of range queries on the database. Study of DBSCAN has few important inclusions too like OPTICS, R-Tree Index and Single Linkage Clustering. A detailed research and study on these methods will be carried out subsequently. Mean-Shift is another method of Density Based Clustering in which using the Kernel Density Estimation, the object is moved to the nearest densest area, eventually, reaching to a local maxima. These local maxima can be the representative of the database.

CLUSTERING APPROACH

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Clustering can be considered the most important unsupervised learning technique. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind. It deals with finding a structure in a collection of unlabelled data. Clustering is the process of organizing objects into groups whose members are similar in some way.[10]

Cluster analysis has been widely used in many applications such as business intelligence image pattern recognition web search biology and security. In business intelligence clustering can be used to organize a large number of customers into groups where customers within a group share similar characteristics. This facilitates the development of business strategies for enhanced customer relationship management. In image recognition clustering can be used to discover cluster or subclasses in handwritten character recognition system. Suppose we have a data set of handwritten digits where each digit is labelled as either 1,2,3, and so on. Note that there can be a large variance in the way in which people write the same digit. Take the number 2, for example, some people may write it with a small circle at the left bottom part, while some other may not. We can use clustering to determine sub classes for each of which represents a variation on the way in which 2 can be written. Using multiple models based on the subclasses can improve overall recognition accuracy.[11]

Clustering in Agriculture

Database of agriculture is increasing day by day. Though, the application of various data mining techniques in the field of agriculture is still relatively new. In this paper we present a brief review of a various of Data Mining techniques that have been applied in the agricultural field. Though limited, data mining techniques like Fuzzy C-Means, Support Vector Machine (SVM), K nearest Neighbour, K-Means, Naïve Bayes Classifier, Neural Networks (NN) and bi-Clustering. How appropriate the data mining technique is, is determined, to some extent, by the problem which is being worked upon or the types of different agricultural data. Few surveys summarize the application of data mining techniques and predictive modelling application in the agriculture field.[12]

Agriculture is the backbone of our country’s economy. As Mahatma Gandhi said, "India lives in villages and agriculture is the soul of Indian economy"[13]

A country’s population survives on the food it grows and that makes a country independent or dependent. With India being an Agricultural economy, India is one of the largest producers of various agricultural products and hence can be called as the “*thali of the world*”, Cuba can be called as Sugar Bowl of the world only based on it’s sugar production.

India has been an agricultural economy since the very beginning. India, unlike many other countries, has almost 70% of its population dependent on Agriculture. Crop production is dependent on various factors including weather, rain, soil, pesticides, fertilizers amongst others. Few researchers have used soil as a parameter to increase crop production in India[14].In this paper, it was stated that the common problem in farmers, in India, is they measure approximate amount of fertilizers and add to the soil for maintaining nutrient levels in soil, in case of deficiency. A wrong composition can harm the soil and the crop. This research provides review of various Data Mining techniques on soil dataset for fertilizer recommendation. Focus has been kept on soil parameter values like of Fe, S, Zn, Cu, N and Ph etc. Though crop cultivation, price prediction, market analysis, rainfall prediction etc are important parameters, but classification of soil is one of the most important since it provides substantial contribution to the support of farmers. Researchers found that data mining in soil datasets is modern and exciting research area.

Many researchers have contributed their knowledge in data mining for Agricultural sector. There are many simultaneous models available for crop productivity predictions. Since the outcome is dependent on economical and environmental parameters, so, these models can not be used for areas outside agriculture. D. Ramesh, B Vishnu Vardhan compared statistical model Multiple Linear Regression over the Data Mining Density-based clustering technique. They concluded that comparison of the crop yield prediction can be made with the entire set of existing available data and will be dedicated to suitable approaches for improving the efficiency of the proposed technique [15, 16 &17].

Usage of data mining techniques in Agriculture is continuing area of research. These techniques can be used to achieve the ultimate goal of increasing the yield of the agricultural sector. Researchers have proposed to get optimized result from the agricultural datasets by using data mining, optimizing techniques on soil composition datasets to be able to give proper recommendations of crop on the basis of soil condition and other factors[18].It can also help in suggesting the right and right mix of fertilizers and pesticides for the recommended crop. Such analysis can also be extended to help farmers improve their economic status by advising them side business during lean periods. These researchers proposed a two stage model. In first stage they applied association rule mining on the agriculture historical data and generated rules from frequent item sets by applying the proper support and confidence for each rule. The user then gave a minimum support and confidence and based on this initial best rule that forms the initial population for GA is extracted. In the second stage, researchers apply Genetic algorithm to optimize the initial population rules which is received from association rule mining. This gives best rules that predict output as an optimized agriculture crop. This research talks about efficient data mining algorithms for agriculture datasets. These datasets use different classification and clustering data mining techniques to predict and group data respectively. In the area of classification for prediction, Logical Model Tree (LMT) classification algorithm is the best fit compared to other algorithms performance. While for clustering, K-Means gives the best results as compared to other clustering algorithms.

Some of the other researches carried out on the agricultural datasets, researchers used Naïve Bayes data mining technique to classify soils for analysing large soil profile experimental datasets. In a research paper[19], researchers used Decision Tree algorithm for predicting soil fertility. Researchers also used clustering techniques of Partitioning & Hierarchical algorithms to determine the land utilization for agriculture and non-agriculture areas for the past ten years.

In another research, the researches tabulated the various applications of various data mining techniques in the agricultural field, as under[20]:

Technique Name	Application in Agriculture
Decision Tree Analysis	Prediction of Soil
K-means	Forecasting Pollution, Combined Classification of Soil with GPS.
K-nearest Neighbour	Simulating precipitations and other weather variable daily
Support Vector Machine	Analysing various changes to weather
Unsupervised Clustering	Generating clusters and finding any pattern
WEKA Tool	Classification system for sorting and grading mushrooms

Data mining process results in discovering new patterns in large data sets. It is the process of analyzing data from different perspectives and summarizing it into useful information without any restriction to the type of data that can be analyzed. The goal of the data mining process is to extract knowledge from an existing data set and transform it for advanced use.

The data availability can be as either a text file or a web server log or a data warehouse or a relational database. Effective Analysis of data in needs understanding of appropriate techniques of data mining. The intention of this paper is to discuss different data mining techniques in perspective of agriculture domain with respect to Quality assessment of Groundnuts so as to develop a model for application of the model to Quality decision of procurement of any of the 16 notified agricultural commodities like Oilseeds, Pulses and Cotton in accordance with Fair Average Quality Norms.

Classification, Clustering and Regression can also be used for accessing Quality and making apt decisions based on the Quality norms specified for agricultural commodities like groundnuts. These technique scan be used to develop suitable data models to achieve high precision and high generality with respect to data points like Maximum limits of tolerance of Foreign Matter (like dust, dirt, stones, lumps of earth, chaff, stem/straw or any other impurity), Damaged Pods, Shrivelled & Immature pods, Pods of other variety, Shelling (kernels/pods) and Moisture Content in the case of groundnut crop.

Data mining process results in discovering new patterns in large data sets. It is the process of analyzing data from different perspectives and summarizing it into useful information without any restriction to the type of data that can be analyzed. The goal of the data mining process is to extract knowledge from an existing data set and transform it for advanced use.

The data availability can be as either a text file or a web server log or a data warehouse or a relational database. Effective Analysis of data in needs understanding of appropriate techniques of data mining. The intention of this paper is to use different data mining techniques in perspective of agriculture domain w.r.t. Quality assessment of Groundnuts so as to develop a model for application of the model to Quality decision of procurement of any of the 16 notified agricultural commodities like Oilseeds, Pulses and Cotton in accordance with Fair Average Quality Norms.

Clustering Techniques can be used for accessing Quality and making apt decisions based on the Quality norms specified for agricultural commodities. Useful research can be done using suitable data models to achieve high precision and high generality w.r.t. to data points like Maximum limits of tolerance of Foreign Matter (like dust, dirt, stones, lumps of earth, chaff, stem/straw or any other impurity), Damaged Pods, Shrivelled & Immature pods, Pods of other variety, Shelling (kernels/pods) and Moisture Content.

SUMMARY

In a country like India, the economy is a lot influenced by Agricultural sector. The success or failure of Agricultural sector is dependent on the rainfall, climate of all seasons. Though, there is a lot of use of technology in the field of Agriculture, usage of Data Mining Techniques in the Agricultural sector of India is still minimal. When it specially comes of usage of humungous data to draw predictive models, very few researches have been done till now. Some of good application of Data Mining Techniques have been seen in studies done by Saeed Soltani and Reza Modarres. Still, the field of Data Mining can be termed as relatively unexplored, especially in Indian context.

Today, where Data Mining has gained momentum across various Industries, Agriculture field must also work on it diligently. We believe that Data Mining techniques like Agglomerative Clustering, DBSCAN, EM Algorithms, K-Means will bring in an advancement that Agricultural sector has long been waiting for.

Devising of models to help procuring the right quality of seeds, would go a long way in shortening the process time between Growing in fields to the shelves of the retail stores. Such models will also help a lot, in a way, to cut down the costs involved with warehousing especially on storage costs, losses because of rotting etc. Though lot has achieved with the application of cluster analysis but still there are many areas untouched and lot of efforts are required to achieve meaningful information.

REFERENCES

[1] Amandeep Kaur & Navneet Kaur: Clustering Techniques

[2] Shraddha K. Popat et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 805-812

[3] Developing innovative applications in agriculture using data mining- Sally Jo Cunningham and Geoffrey Holmes

[4] Tanuj Wala: International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), May 2015

[5] Everitt, Brian (2011). Cluster analysis. Chichester, West Sussex, U.K

[6] Lloyd, S. (1982). "Least squares quantization in PCM".

[7] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm"

[8] Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering".

[9] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise".

[10] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Dorling Kindersley Pvt. Ltd. India, Sixth Edition,2013.

[11] Tayel, Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141- 149,2014

[12] Dharmendra Patel: International Journal of Computer Applications." A Brief survey of Data Mining Techniques Applied to Agricultural Data"

[13] http://www.klientsolutech.com/agriculture-in-india/

[14] A Review on Data Mining Techniques for Fertilizer Recommendation, Jignasha M. Jethva, Nikhil Gondaliya, Vinita Shah

[15] Review of Literature of Data Mining Techniques for Crop Yield Prediction shital H. Bhojani-2017

[16] Data mining Concepts and Techniques by Jiawei Han and Micheline Kambe

[17] Analysis of Data Mining Techniques for Agriculture Data, E.Manjula*, S. Djodiltachoumy

[18] Mr.Omkar B. Bhalerao and Prof. L. M. R. J. Lobo

[19] Efficient Data Mining Algorithms For Agriculture Data Anusha A. Shettar, Shanmukhappa A. Angadi

[20] Megala, S., & Hemalatha, M. (2011). A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu. International Journal of Computer Applications

[21] https://www.thehindubusinessline.com/markets/commodities/gujarat-nafed-to-procure-groundnut-jointly/article25540886.ece